Machine Learning Engineer Nanodegree

Capstone Proposal – Dogs vs. Cats

领域背景

猫狗大战（Dogs vs. Cats）项目所属的领域是计算机视觉（Computer Vision），它可以简单的概括为“用计算机代替人眼，在图片中重建和解释世界”。1982 年 David Marr 发表《Vision》，提出了计算机视觉的表达、算法和硬件实现三个层次，奠定了该领域的研究格局。

本项目所涉及的图像分类问题（Image Classification）是目标识别（Object Recognition）任务中比较基础且重要的方向，基于此而开发的成果可应用于目标检测或图像摘要生成等其他方向。

近年来，随着数据量和计算能力的显著提升，连接注意所倡导的人工神经网络（ANN）在几经兴衰之后更名为深度学习重新回到人们的视野并成为焦点。2012 年开始，深度学习方法在 ImageNet 挑战赛中独领风骚。事实证明，深度学习在图像分类、目标识别等等计算机视觉领域有着非常好的表现。

问题描述

项目所要解决的图像分类，对应于监督学习中的二分类问题（猫和狗）。具体来讲，就是选择适当的机器学习模型并使用带标签的数据集来训练该模型（求解参数），完成训练后将模型泛化至未知数据，即对未知图像进行分类。解决此问题的方法可以是传统机器学习中的支持向量机（SVM）、AdaBoost 以及近来非常热门的深度学习（Deep Learning）。

从数据角度看，输入图片是一个包含许多像素值的矩阵或者向量，经过模型运算输出某个特定类别（比如狗）的概率，因此该问题是可量化的（quantifiable）；机器学习的目的是使模型输出特定类别的概率尽可能接近 1（真实类别）或 0（非真实类别），因此该问题是可衡量的（measurable）；另外，对于任意一个输入图片，模型总能将输入映射为一个概率值，因此该问题是可复制的（replicable）。

输入数据

输入数据全部来自于 kaggle，可以从网站上进行下载。其中包含训练数据和测试数据两部分，全都是各种猫或狗的图片，因此对本项目是适用的。其中，训练数据将会被分为两部分，其中八成左右用于训练模型，二成左右用于验证模型表现并作为调整参数的依据；测试数据用于评估训练完成的模型的泛化能力。上述对训练集和测试集的使用，符合监督学习的一般流程，对该项目也是适合的。

训练数据共有 25000 张图片，猫和狗各占一半，每张图片都带有类别标签。由于猫狗的比例为 1:1，因此不需要考虑类别不平衡问题。测试数据共有 12500 张图片。在上述所有图片中，分辨大小各异，但全都包含 RGB 3 个 channel 的信息。

解决办法

拟通过深度学习中的卷积神经网络（CNN）方法来解决该问题。CNN 是一种专门用来处理具有类似网格结构的数据（图片即是如此）的神经网络，因此将 CNN 用于解决图像识别问题是很合适的。CNN 网络将输入数据进行若干卷积层以及全联接层处理，在输出层给出两个节点并进行 softmax 计算得到两个类别各自的概率，因此该方案是可量化的（quantifiable）；定义交叉熵（Cross Entropy）为模型的损失函数，因此该方案是可衡量的（measurable）；基于梯度下降算法（Gradient Descent）配合特定的优化（如 Mini-Batch，Momentum，正则化等）理论上可以找到参数的最优解（或次优解），整个优化过程是可复制的（replicable）。

基准模型

近年来，在 ImageNet 挑战中赛涌现了许多优秀的卷积神经网络架构，如 Inception V3，ResNet 等。前者通过 Inception 的堆叠使整个网络的宽度和深度都可以扩大，相应带来计算效能和性能的提升；后者通过残差块（Residual Block）的堆叠，可令网络达到非常深的深度（152-layer）同时具有非常好的性能。ResNet 在 2015 年以 Top-5 错误率 3.6%横扫其他对手拿到图像分类比赛第一名。

幸运的是，ImageNet 猫和狗的分类，上述网络也有公开训练好的模型。因此，可直接将这些模型作为基准模型，可与本项目的方案做客观对比。基准模型的输出也是以 softmax 计算特定类别的概率，与本项目的方案一致，同样也是可衡量的（measurable）。

评估指标

评估指标采用对数损失（LogLoss）来衡量：

$$LogLoss = -\frac{1}{n}\sum_{i=0}^{n}[y_i \log(\hat{y_i}) + (1-y_i)\log(1-\hat{y_i})]$$

其中：

- n 是图片的数量
- $\hat{y_i}$是模型预测为狗的概率
- y_i是类别标签，1 对应狗，0 对应猫
- log ()是自然对数

可见，对数损失越小，代表模型的性能越好。上述评估指标可用于评估该项目的解决方案以及基准模型。

设计大纲

该项目的设计主要包含以下几个部分：

- 数据预处理
- 选择并设计网络架构
- 选择并实施优化算法训练网络
- 根据验证集的表现调整参数

1. 在数据预处理阶段，由于图片的分辨率大小不一，因此要根据模型的输入需求统一调整图片的大小。此外，对数据的处理还包括零中心化（zero-centered）或者是[-1, 1]区间的归一化。

2. 网络结构选择和设计是整个方案中的重要一环。可以尝试先堆叠一些卷积层/池化层/Batch Normalization，查看训练的效果并调整相应的网络结构和参数。

   如果上述做法效果不是很好，可考虑采用迁移学习（Transfer Learning）方法，即使用已有的训练好的 CNN 模型。如前述，目前已经存在于实战中性能表现非常好的模型，站在前人的基础上再做改进也许是一个不错的选择。基于迁移学习也有两种具体的做法：

- 其一是直接使用基础模型进行少量参数的微调（fine-tune），对于样本量较小的数据集是一个不错的选择。
- 其二是将基础模型作为特征提取器，再设计一个分类器模型，进行更多参数的调整。比较适合拥有多一些数据量的数据集。

3. 在优化算法的选择上，可以是 Mini-Batch SGD，或者是学习率自适应算法如 Adagrad/RMSprop/Adam，对于训练比较深比较复杂的网络，倾向于使用学习率自适应算法。

4. 在训练过程中观察模型在验证数据上的表现(依据前述的评估指标)，调整 Batch Size/Epoch 等超参数。

上述步骤可反复交叉进行直至评估指标达到较好成绩。